Chemical and genomic evolution of enzyme-catalyzed reaction networks.

AUTHOR(S):

Kanehisa, Minoru

CITATION:

Kanehisa, Minoru. Chemical and genomic evolution of enzyme-catalyzed reaction networks.. FEBS letters 2013, 587(17): 2731-2737

ISSUE DATE:

2013-09-02

URL:

http://hdl.handle.net/2433/178762

Chemical and genomic evolution of enzyme-catalyzed reaction networks

Minoru Kanehisa

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Tel: +81-774-38-4521

Fax: +81-774-38-3269

E-mail: kanehisa@kuicr.kyoto-u.ac.jp

ABSTRACT

There is a tendency that a unit of enzyme genes in an operon-like structure in the prokaryotic genome encodes enzymes that catalyze a series of consecutive reactions in the metabolic pathway. Our recent analysis shows that this and other genomic units correspond to chemical units reflecting chemical logic of organic reactions. From all known metabolic pathways in the KEGG database we identified chemical units, called reaction modules, as the conserved sequences of chemical structure transformation patterns of small molecules. The extracted patterns suggest co-evolution of genomic units and chemical units. While the core of the metabolic network may have evolved with mechanisms involving individual enzymes and reactions, its extensions may have been driven by modular units of enzymes and reactions.

KEYWORDS: evolution of metabolism; metabolic pathway; enzyme cluster; reaction module; KEGG database.

1. Introduction

Leonor Michaelis was a professor of Aichi Medical College (currently Nagoya University School of Medicine) in Japan from late 1922 to early 1926. His name associated with enzyme kinetics [1] is well recognized, but the fact that the actual person spent three years in Nagoya is no longer widely known among Japanese scientists. Nevertheless, this review is a tribute to his presence in Japan, especially to his contribution to the early days of Japanese biochemistry [2]. In 1995 we started the KEGG (Kyoto Encyclopedia of Genes and Genomes) database project under the then ongoing Human Genome Program in Japan. The original concept was to create a reference knowledge base of metabolism and other cellular processes from published literature, so that it can be used for biological interpretation of genome sequence data. The KEGG database has expanded significantly over the years to meet the needs for integrating and interpreting large-scale datasets generated by various types of high-throughput experimental technologies [3], but this basic concept is unchanged. At first the KEGG metabolic pathway maps were created using the book “Metabolic Maps” [4] compiled by the Japanese Biochemical Society. This Society was founded in 1925 during Michaelis’ stay in Japan, and the biochemistry of enzymes was an active field since then. The original KEGG that owes to this tradition still remains in the metabolic pathway section of the KEGG PATHWAY database. The KEGG pathway map identifiers such as map00010, map00020, and map00030 for glycolysis, citrate cycle, and pentose phosphate pathway correspond to the map numbers 1, 2, and 3 in the Japanese Biochemical Society’s Metabolic Maps.

Since its inception the KEGG metabolic pathway map is drawn to represent two types of networks: the chemical network of how small molecules are converted and the genomic network of how genome-encoded enzymes are connected to catalyze consecutive reactions. This dual aspect has been utilized for metabolic reconstruction.

A set of enzyme genes encoded in the completely sequenced genome will identify enzyme relation networks when superimposed on the KEGG pathway maps, which in turn characterize chemical structure transformation networks allowing interpretation of biosynthetic and biodegradation potentials of the organism. In addition to this type of genome analysis, the KEGG metabolic pathway maps can be used for chemical analysis of small molecules and reactions [5-8]. This review focuses on our efforts to integrate genomics and chemistry toward better understanding of intrinsically related genome evolution and chemical evolution of enzyme-catalyzed reactions.

2. The KEGG resource

2.1. KEGG metabolic pathway map

The KEGG metabolic pathway maps are graphical diagrams representing knowledge of enzyme-catalyzed reaction networks. Each map is manually drawn to capture the overall architecture of how main compounds are converted. The details of individual reactions involving all substrates and products can be examined in the KEGG REACTION entries linked from the map. It is also drawn as a generic map combining and summarizing experimental evidence in different organisms, so that it can be used for interpretation of any genome. This is accomplished by the KEGG Orthology (KO) system described below. Basic graphics objects in the KEGG metabolic pathway maps are boxes for enzymes and circles for chemical compounds (see, for example, http://www.kegg.jp/pathway/map00010). Each circle is identified by the chemical compound identifier (C number). Each box is given two types of identifiers: the reaction identifier (R number) and one or more KO identifiers (K numbers). Although the EC (Enzyme Commission) numbers are usually displayed in the boxes, they are not identifiers and are treated as attributes to KO identifiers. Note that the EC numbers may represent reaction classification of the EC system or gene/protein functional

classification in the genome annotation. These two aspects of enzymes are clearly separated by the R number and the K number identifiers in KEGG, enabling the analysis of chemical networks and genomic networks in much better defined ways than using the EC numbers.

2.2. KEGG Orthology

The KEGG Orthology (KO) system is a collection of manually defined ortholog groups (KO entries) for all proteins and functional RNAs that appear in the KEGG pathway maps (both metabolic and non-metabolic) as well as in the KEGG BRITE functional hierarchies (ontologies). Whenever a pathway map is drawn based on experimental observations in specific organisms, an additional manual work is performed for generalizing gene information from those specific organisms to other organisms. This is done by assigning KO entries to the map objects (boxes) and, when necessary, by defining a new KO entry and creating a corresponding set of orthologous genes from available genomes. Each KO entry also represents a sequence similarity group. This allows computational assignment of KO identifiers in newly determined genomes and metagenomes by sequence comparison, which may then be used for KEGG pathway mapping (reconstruction) analysis. Note that the degree of similarity in each group varies significantly because each KO is defined in a context (pathway) dependent manner.

2.3. KEGG reaction class

The KEGG REACTION database contains all biochemical reactions that appear in the KEGG metabolic pathway maps together with the set of experimentally characterized enzymatic reactions in the Enzyme Nomenclature [9], i.e., those with the official EC numbers. Less than one half of the reactions in the KEGG pathway maps

correspond to the Enzyme Nomenclature reactions, suggesting the difficulty of using EC numbers for a comprehensive analysis. In order to analyze chemical compound structure transformation patterns, the following processing is performed for all reactions both computationally and manually. First, reactant pairs are defined as one-to-one relationships of substrate-product pairs by considering the reaction type (as classified by the EC system) and the flow of atoms. Second, structure transformation patterns are computed, manually curated, and represented by the so-called RDM patterns of KEGG atom type changes [5-7]. Third, the identity of RDM patterns for the main reactant pairs, i.e. the reactant pairs that appear in the KEGG pathway maps, is used to define KEGG reaction class [8]. The resulting KEGG reaction class (identified by RC number) is like an ortholog group of reactions defined by localized structural changes and accommodating global structural differences of reactants.

2.4. KEGG module

Functional units of enzyme complexes and subpathways are often encoded in positionally correlated gene sets (operon structures) in prokaryotic genomes. When complete genome sequences first became available, a graph analytical method was used to extract enzyme gene clusters on the chromosome that encode consecutive reaction steps in the metabolic pathways [10]. Such functional units are now accumulated in the pathway module section of the KEGG MODULE database. Each KEGG module (identified by M number) is manually defined as a combination of KO identifiers. For example, the reaction sequence involving oxaloacetate + acetyl-CoA, citrate, isocitrate, and 2-oxoglutarate in the citrate cycle (map00020) is the KEGG pathway module M00010 named as “Citrate cycle, first carbon oxidation” and defined by:

K01647 (K01681,K01682) (K00031,K00030)

where alternative enzymes are given in parentheses. The positional correlation of operon-like structures is not always observed, but when it exists, at least, in certain organism groups, as is the case for many KEGG pathway modules, it well supports the definition of functional units.

2.5. KEGG reaction module

An alternative way to define functional units in the metabolic pathways has been developed recently [8]. It relies only on the chemistry of reactions without using the information about genes and proteins. As mentioned, KEGG pathway nodes (boxes) are given both K numbers (gene/protein orthologs) and R numbers (reactions), where the latter can be converted to RC numbers (reaction class or reaction orthologs). While KEGG pathway modules are conserved subnetwoks of the K number network, different types of conserved subnetworks may exist in the RC number network. This is in fact the case, and conserved reaction sequences termed reaction modules can be extracted from known metabolic pathways [8]. Furthermore, reaction modules (also called RC modules) tend to correspond to KEGG pathway modules (also called KO modules) despite the fact that they are separately defined from different properties. A case in point is the RC module RM001, which exactly matches the KO module M00010, for the reaction sequence from oxaloacetate to 2-oxoglutarate. RM001 is named as “2-Oxocarboxylic acid chain extension by tricarboxylic acid pathway” and defined by:

RC00067 (RC00498+RC00618,RC00497) (RC00084+RC00626,RC00114)

RC01205 RC00976+RC00977 RC00417

RC00470 RC01041+RC01046 RC00084+RC00577

where the notation is somewhat more complex because of the existence of three subtypes and multi-step reactions denoted by plus signs.

3. Modular architecture of metabolic network

3.1. Reaction modules used in combination

The analysis of reaction modules has revealed the modular architecture of the metabolic network with two interesting aspects: the existence of chemical units containing chemical logic of organic reactions and the correspondence of chemical and genomic units [8]. The chemical units of reaction modules are used in combination as if they are building blocks of the metabolic network, generating different chemical substances in different pathways. A notable example is illustrated in Figure 1 for 2-oxocarboxylic acid chain elongation and modification as part of the biosynthesis pathways of branched-chain amino acids (valine, leucine, and isoleucine) and basic amino acids (arginine and lysine). 2-Oxocarboxylic acids are an important class of precursor metabolites including pyruvate (monocarboxylic acid), oxaloacetate (dicarboxylic acid), and 2-oxoisovalerate (methyl-modified monocarboxylic acid). The increase of the chain length from these three is shown by vertical arrows in Figure 1, where eight 2-oxocarboxylic acids are denoted by shaded (red) circles. All the vertical arrows correspond to the reaction module RM001 for increasing the 2-oxocarboxylic acid chain length by one using acety-CoA as a carbon source. This module consists of a series of characteristic reactions involving tricarboxylic acids.

Each of the six 2-oxocarboxylic acids, excluding pyruvate and 2-oxobutanoate, is followed by a single reaction step of reductive amination (RC00006 or RC00036) indicated by a horizontal line. The reaction modules for 2-oxocarboxylic acid chain modifications are shown by horizontal arrows. RM033 is for branched-chain addition and RM032 and RM002 are for carboxyl to amino conversion. There is an interesting distinction between RM032 and RM002 defined by:

RM032: RC00043 RC00684 RC00062

RM002: RC00064 RC00043 RC00684 RC00062 RC00064

RM032 for a shorter chain is a direct conversion via a phosphorylation step (RC00043), but RM002 uses a protective N-acetyl group that is added before and removed after (both RC00064) the core sequence of RM032. Furthermore, for a longer chain in lysine biosynthesis, the protective N-acetyl group is attached to a carrier protein.

A similar variation exists in the biosynthetic pathways of fatty acids, in which the acyl chain length is increased by two in one cycle of the four-step reaction sequence consisting of ketoacyl synthase (KS), ketoreductase (KR), dehydratase (DH), and enoylreductase (ER) reactions. Two slightly different reaction modules are identified for this sequence: RM021 for the major pathway and RM020 for the minor pathway in mitochondria. The minor pathway, which is essentially the reversal of beta oxidation, does not involve acyl carrier protein and uses acetyl-CoA as a carbon source. In contrast, the major pathway, which may be a more recent invention, involves acyl carrier protein and uses malonyl-CoA as a carbon source. These examples suggest that the reaction modules seem to contain design principles of a series of organic reactions including how to achieve an activated transition state (e.g., phosphorylation), how to introduce a protective group (e.g., N-acetylation), and how to increase specificity and efficiency (e.g., using carrier protein and switching from acetyl-CoA to malonyl-CoA). This type of chemical logic may have played roles in the chemical evolution of metabolic networks.

3.2. Correspondence of chemical and genomic units

Reaction modules are derived from purely chemical properties of substrate-product structure transformation patterns, but they are found to correspond to KEGG pathway modules defined as sets of enzyme orthologs in the genome. This is already mentioned for the correspondence of RM001 and M00010. In fact, as shown in Figure 1 the chain elongation module RM001 corresponds to different enzyme units in different pathways:

M00010 for oxaloacetate to 2-oxoglutarate in citrate cycle, M00433 for 2-oxoglutarte to 2-oxoadipate in lysine biosynthesis, M00535 for pyruvate to 2-oxobuanoate in isoleucine biosynthesis, and M00432 for 2-oxovalerate to 2-oxosicaproate in leucine biosynthesis. There are two characteristic features. First, the enzyme units are encoded in operon-like gene clusters, at least, in certain genomes. Second, the enzyme units are similar in the sense that they contain paralogous genes coding for similar amino acid sequences.

These features are commonly observed in the KEGG MODULE database. Here a few examples of RM001 are shown in Table 1 for the genomes of *Lactococcus lactis* IO-1 [11], *Bacteroides fragilis* YCH46 [12], *Thermus thermophilus* HB27 [13] and *Saccharomyces cerevisiae*. The numbering of gene identifiers (locus tags) in the three prokaryotic genomes indicates the proximity of genes on the chromosome. Among the three enzymes (or enzyme complexes) that constitute RM001, the second dehydratases and the third dehydrogenases clearly form paralogous gene groups with sequence identity ranging 35% to 45%. The first enzymes, citrate synthase, homocitrate synthase and 2-isopropylmalate synthase, are more distantly related; only the enzymes in the lysine and leucine pathways exhibit some sequence similarity. Thus, the same reaction sequence appears to be generated in different pathways by duplicating and slightly modifying enzyme clusters. The enzyme that catalyzes the first step of the reaction sequence tends to be more divergent, possibly reflecting the constraint of specific substrate recognition.

Other notable examples can be found in the microbial biodegradation pathways. Certain groups of bacteria are capable of metabolizing non-biological chemicals accumulated in the environment by acquiring or modifying gene sets for biodegradation. These gene sets are often encoded in plasmids and can be transferred within a bacterial community. Environmental pollutants such as endocrine disrupting compounds

are mostly aromatic compounds, and the modular architecture of the biodegradation pathways consists of well-defined reaction modules for aromatic ring cleavage. They include preprocessing modules of methyl to carboxyl conversion on aromatic ring (RM004 and RM013) and the main modules of dihydroxylation and cleavage. Four types of dihydroxylation modules are defined distinguishing dioxygenase and dehydrogenase reactions (RM004), dioxygenase and decarboxylating dehydrogenase reactions (RM005), two monooxygenase reactions (RM006), and dealkylation and monooxygenase reactions (RM007). Two basic types of cleavage modules are defined for ortho-cleavage (RM008) and meta-cleavage (RM009). Not surprisingly, these reaction modules well correspond to enzyme gene clusters in operon-like structures defined as KEGG pathway modules. For example, the BTX (benzene, toluene, and xylene) degradation capacity is well represented by the corresponding sets of reaction modules and KEGG pathway modules: preprocessing of toluene to benzoate (RM003 and M00538) or xylene to methylbenzoate (RM003 and M00537), dihydroxylation of benzene to catechol (RM006 and M00548), dihydroxylation of benzoate to catechol (RM005 and M00551), meta-cleavage of catechol (RM009 and M00569), and ortho-cleavage of catechol (RM008 and M00568). These observations suggest a link between genomic diversity and chemical diversity. It should be emphasized again that the link is not simply between individual genes and reactions, but rather between genomic units and chemical units reflecting the modular architecture of the metabolic network.

3.3. Degree of modularity

The modular architecture of reaction modules and enzyme gene sets was most apparent in carboxylic acid metabolism (for 2-oxocarboxylic acids and fatty acids) and aromatics degradation. However, such modularity cannot explain the architecture of the

entire metabolic network. Figure 2 is an overview map for the biosynthesis of twenty amino acids (http://www.kegg.jp/pathway/map01230). Circles represent chemical compounds and lines connecting them are reactions (or sets of reactions). The twenty amino acids are shown in shaded (red) circles, the reaction modules RM001 and RM002 are shown in thick (blue) lines, and the KEGG pathway modules are shown as separate (red) lines with M numbers attached. This overall pathway may be viewed as consisting of the core part and its extensions.

The core part is the pathway module M00002 for conversion of three-carbon compounds from glyceraldehyde-3P to pyruvate, together with the pathways around serine and glycine. M00002 is the most conserved pathway module in the KEGG MODULE database and is found in almost all the completely sequenced genomes. The extensions are the pathways containing the reaction modules RM001 and RM002 for biosynthesis of branched-chain amino acids (left) and basic amino acids (bottom), and the pathways for biosynthesis of histidine and aromatic amino acids (top right). Note that no reaction modules are extracted by our method [8] for the biosynthetic pathways of histidine and aromatic amino acids because they are not shared in other pathways. However, they can be considered uniquely defined modular units because of the existence of enzyme gene clusters. It is interesting to note that the so-called essential amino acids that cannot be synthesized in human and other organisms generally appear in these extensions. Furthermore, the bottom extension of basic amino acids appears to be most divergent containing multiple pathways for lysine biosynthesis and multiple gene sets for arginine biosynthesis.

Figure 2 shows only the pathways that are relevant to amino acid biosynthesis. What constitutes the core part of the entire metabolic network and how it has evolved would require more detailed analyses of the central energy metabolism in relation to diverse environmental conditions in which various organisms inhabit. The increasing

amount of genome sequences and metagenome sequences, together with the accumulated knowledge of metabolism as represented in KEGG pathway maps, will enable such analyses to be performed.

4. On the evolution of metabolic networks

The idea of conserved core and divergent extensions in the metabolic network is hardly new. The distinction of primary and secondary metabolism contains a similar notion. The core is required for maintaining life and is conserved among all organisms. The extensions are required for interactions with the environment and are specific to certain organism groups. Microbial biodegradation pathways are typical examples of secondary metabolism, converting xenobiotic compounds with varying structures into a limited number of compounds in primary metabolism. Here we assert that even within primary metabolism there is a primitive core and its extensions. The conversion of three-carbon compounds from glyceraldehyde-3P to pyruvate (M00002) is followed by the first segment of citrate cycle from oxaloacetate to 2-oxoglutarate (M00010). We view M00002 as part of the primitive core and M00010 as a modular extension as illustrated in Figure 2. According to the genome annotation in KEGG these two modules are highly conserved, but there are certain organisms that apparently lack M00010 [14]. In contrast, the second segment of citrate cycle from 2-oxoglutarate to oxaloacetate (M00011) exists only in less than one half of the completely sequenced genomes. Citrate cycle may thus be an invention of combining one ancient module and another more recent module.

Here we investigate more on the central carbon metabolism illustrated in Figure 3. The number of carbons is shown for each compound denoted by a circle, excluding CoA or THF (tetrahydrofolate), which is replaced by an asterisk. This map is first drawn by combining the three KEGG pathway maps, Glycolysis (map00010), Pentose phosphate

pathway (map00020), and Citrate cycle (map00030), whose reaction steps are denoted by blue arrows and which are here called carbon utilization pathways. Then, by using two more KEGG maps, Carbon fixation in photosynthetic organisms (map00710) and Carbon fixation pathways in prokaryotes (map00720), six known carbon fixation pathways [15,16] are superimposed. They are: (1) reductive pentose phosphate cycle (Calvin cycle) in plants and cyanobacteria that perform oxygenic photosynthesis, (2) reductive citrate cycle in photosynthetic green sulfur bacteria and some chemolithoautotrophs, (3) 3-hydroxypropionate bi-cycle in photosynthetic green nonsulfur bacteria, two variants of 4-hydroxybutyrate pathways in Crenarchaeota called (4) hydroxypropionate-hydroxybutyrate cycle and (5) dicarboxylate-hydroxybutyrate cycle, and (6) reductive acetyl-CoA pathway in methanogenic bacteria. In Figure 3 pathways 1 to 5 are denoted by red arrows and pathway 6 by green arrows.

The differences of these carbon fixation pathways from the utilization pathways can be classified into three types. First, the carbon fixation pathway is a minor variation containing reaction steps catalyzed by key enzymes. This is most apparent in reductive pentose phosphate cycle (pathway 1) shown on top right of Figure 3, in which two additional reaction steps are catalyzed by key enzymes RuBisCO (ribulose-bisphosphate carboxylase) and PRK (phosphoribulokinase). Reductive citrate cycle (pathway 2) on bottom left also belongs to this minor variation type.

Second, the carbon fixation pathway consists of four units of reaction sequences in carbon metabolism: (A) succinyl-CoA to acetyl-CoA roughly corresponding to the second segment of citrate cycle, (B) acetyl-CoA to propionyl-CoA to succinyl-CoA containing three-carbon reaction sequence found in propanoate metabolism (map00640), (C) succinyl-CoA to acetoacetyl-CoA to acetyl-CoA containing four-carbon reaction sequence found in butanoate metabolism (map00650), and (D) propionyl-CoA to acetyl-CoA containing five-carbon reaction sequence. The three overlapping carbon

fixation pathways are formed by these segments: 3-hydroxypropionate bi-cycle (pathway 3 consisting of A, B, and D), hydroxypropionate-hydroxybutyrate cycle (pathway 4 consisting of B and C), and dicarboxylate-hydroxybutyrate cycle (pathway 5 consisting of A and C).

Third, carbon fixation results from a different pathway, methane metabolism, in the case of reductive acetyl-CoA pathway (pathway 6). This pathway appears to represent a most primitive form of carbon fixation. All the other carbon fixation pathways are modifications of existing pathways, whether the modification is incremental (individual reactions and individual enzyme) or modular (units of reactions and units of enzymes). We postulate that a primitive core may exist around reductive acetyl-CoA pathway together with parts of the pathways for methane metabolism (map00680), nitrogen metabolism (map00910), and sulfur metabolism (map00920).

Many models have been presented for the metabolic pathway evolution including the retrograde model [17] and the patchwork model [18,19]. Our analysis indicates an additional aspect; namely, chemical evolution driven by chemical logic of a series of organic reactions. The increasing complexity of the molecular machinery, such as fatty acid synthase, is a genome evolution, but it also reflects the increasing complexity of organic reactions, a chemical evolution. It is unlikely that a single model can explain all aspects of the metabolic pathway evolution. An integrated approach of genomics and chemistry will better characterize the intrinsically related genome evolution and chemical evolution of the metabolic network under the changing environment of Earth.

ACKNOWLEDGEMENTS

This work was partially supported by the Japan Science and Technology Agency. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Michaelis, M. and Menten, M.L. (1913) Die Kinetik der Invertinwerkung. Biochem. Z. 49, 333-369.

[2] Slater, E.C. (2006) Leonor Michaelis in Japan. IUBMB Life 58, 376-377.

[3] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109-D114.

[4] Nishizuka, Y., ed. (1980) Metabolic Maps (in Japanese), Tokyo Kagaku Dojin, Tokyo.

[5] Hattori, M., Okuno, Y., Goto, S. and Kanehisa, M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc. 125, 11853-11865.

[6] Kotera, M., Okuno, Y., Hattori, M., Goto, S. and Kanehisa, M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J. Am. Chem. Soc. 126, 16487-16498.

[7] Oh, M., Yamada, T., Hattori, M., Goto, S. and Kanehisa, M. (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. J. Chem. Inf. Model. 47, 1702-1712.

[8] Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S. and Kanehisa, M. (2013) Modular architecture of metabolic pathways revealed by conserved sequences of reactions. J. Chem. Inf. Model. 53, 613-622.

[9] McDonald, A.G., Boyce, S. and Tipton, K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res. 37, D593–D597.

[10] Ogata, H., Fujibuchi, W., Goto, S. and Kanehisa, M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme

clusters. Nucleic Acids Res. 28, 4021-4028.

[11] Kato, H., Shiwa, Y., Oshima, K., Machii, M., Araya-Kojima, T., Zendo, T., Shimizu-Kadota, M., Hattori, M., Sonomoto, K. and Yoshikawa, H. (2012) Complete genome sequence of *Lactococcus lactis* IO-1, a lactic acid bacterium that utilizes xylose and produces high levels of L-lactic acid. J. Bacteriol. 194, 2102-2103.

[12] Kuwahara, T., Yamashita, A., Hirakawa, H., Nakayama, H., Toh, H., Okada, N., Kuhara, S., Hattori, M., Hayashi, T., Ohnishi, Y. (2004) Genomic analysis of Bacteroides fragilis reveals extensive DNA inversions regulating cell surface adaptation. Proc. Natl. Acad. Sci. 104, 14919-14924.

[13] Henne, A., Brüggemann, H., Raasch, C., Wiezer, A., Hartsch, T., Liesegang, H., Johann, A., Lienard, T., Gohl, O., Martinez-Arias, R., Jacobi, C., Starkuviene, V., Schlenczeck, S., Dencker, S., Huber, R., Klenk, H.P., Kramer, W., Merkl, R., Gottschalk, G. and Fritz, H.J. (2004) The genome sequence of the extreme thermophile Thermus thermophilus. Nat. Biotechnol. 22, 547-553.

[14] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29-34.

[15] Berg, I.A., Kockelkorn, D., Ramos-Vera, W.H., Say, R.F., Zarzycki, J., Hügler, M., Alber, B.E. and Fuchs, G. (2010) Autotrophic carbon fixation in archaea. Nat. Rev. Microbiol. 8, 447-460.

[16] Saini, R., Kapoor, R., Kumar, R., Siddiqi, T.O. and Kumar, A. (2011) CO_2 utilizing microbes – a comprehensive review. Biotechnol. Adv. 29, 949-960.

[17] Horowitz, N.H. (1945) On the evolution of biochemical synthesis. Proc. Natl. Acad. Sci USA, 31, 153-157.

[18] Ycas, M. (1974) On earlier states of the biochemical system. J. Theor. Biol. 44,

145-160.

[19] Jensen, R.A. (1976) Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. 30, 409-425.

Table 1. Examples of enzyme gene clusters for reaction module RM001.

Organism[1]	Gene	M00010	M00433	M00432[2]	Identity[3]
Lactococcus lactis (lls)	synthase	lilo_0538		lilo_1107	–
	dehydratase	lilo_0539		lilo_1109	0.258
				lilo_1110	0.305
	dehydrogenase	lilo_0540		lilo_1108	0.300
Bacteroides fragilis (bfr)	synthase	BF3753		BF3445	–
	dehydratase	BF3755		BF3447	0.255
				BF3446	0.317
	dehydrogenase	BF3754		BF3444	0.250
Thermus thermophilus (tth)	synthase	TTC0978	TTC1550	TTC0849	– / – / 0.307
	dehydratase	TTC0374	TTC1547	TTC0865	0.266 / 0.261
			TTC1546	TTC0866	0.341 / 0.394
	dehydrogenase	TTC1172	TTC1012	TTC0867	0.457 / 0.364
Saccharomyces cerevisiae (sce)	synthase	YNR001C	YDL182W	YNL104C	– / – / 0.222
	dehydratase	YLR304C	YDR234W	YGL009C	0.266 / 0.269
	dehydrogenase	YDL066W	YIL094C	YCL018W	0.322 / 0.280

1, KEGG organism codes in parentheses. For *Saccharomyces cerevisiae* only one set of genes is shown.

2, Multiple genes indicate large and small subunits.

3, Sequence identity between M00010 and M00433 / M00010 and M00432 / M00433 and M00432

Figure Legend

Figure 1. The modular architecture of 2-oxocarboxylic acid metabolism (http://www.kegg.jp/pathway/map01210). The 2-oxocarboxylic acid chain elongation is shown in the vertical direction and its modification in the horizontal direction. The correspondence of reaction modules (RM001, etc.) and KEGG pathway modules (M00010, etc.) is also shown.

Figure 2. An overview map for biosynthesis of amino acids (http://www.kegg.jp/pathway/map01230). The reaction modules of 2-oxocarboxylic acid chain elongation and modification are shown by thick (blue) lines and the KEGG pathway modules (M00002, etc.) are shown by separate (red) lines.

Figure 3. An overview map for central carbon metabolism. The number of carbons is shown for each compound excluding a cofactor. The map combines carbon utilization pathways of glycolysis, citrate cycle, and pentose phosphate pathway (denoted by blue arrows) and six known carbon fixation pathways: reductive pentose phosphate cycle, reductive citrate cycle, 3-hydroxypropionate bi-cycle, hydroxypropionate-hydroxybutyrate cycle, dicarboxylate-hydroxybutyrate cycle (all denoted by red arrows), and reductive acetyl-CoA pathway (denoted by green arrows).

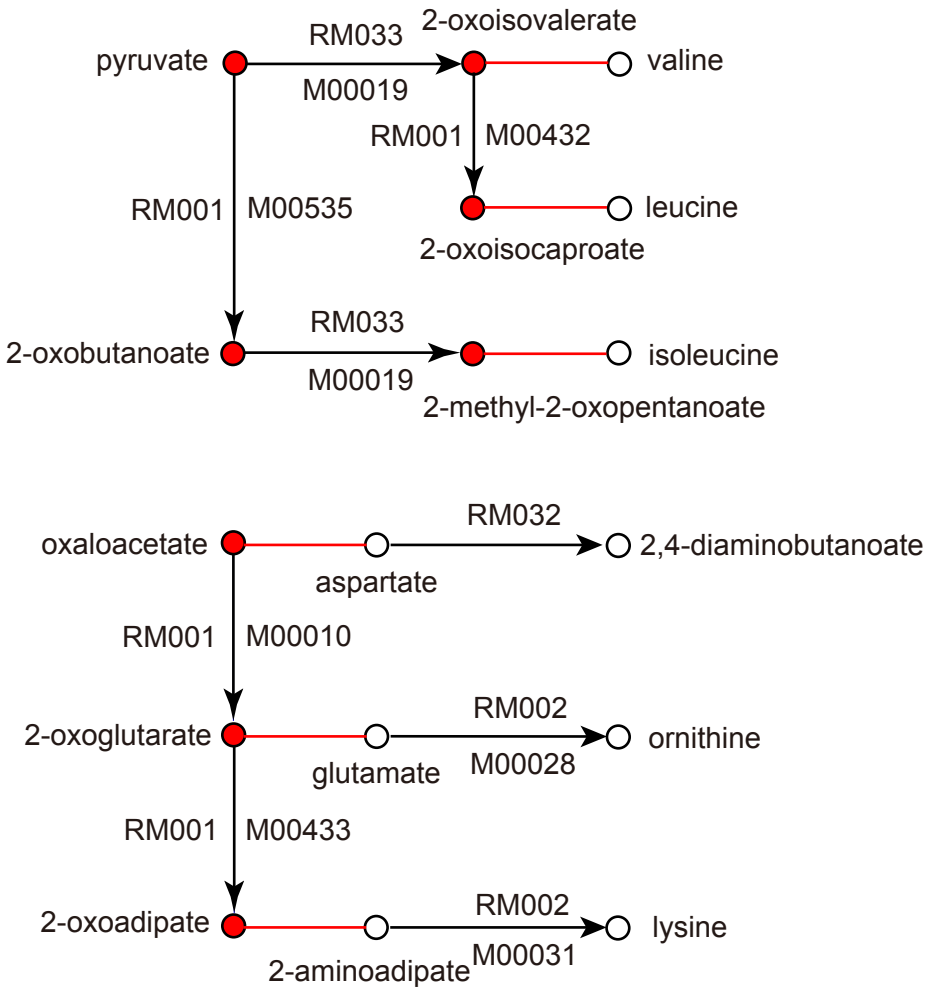

Figure 1

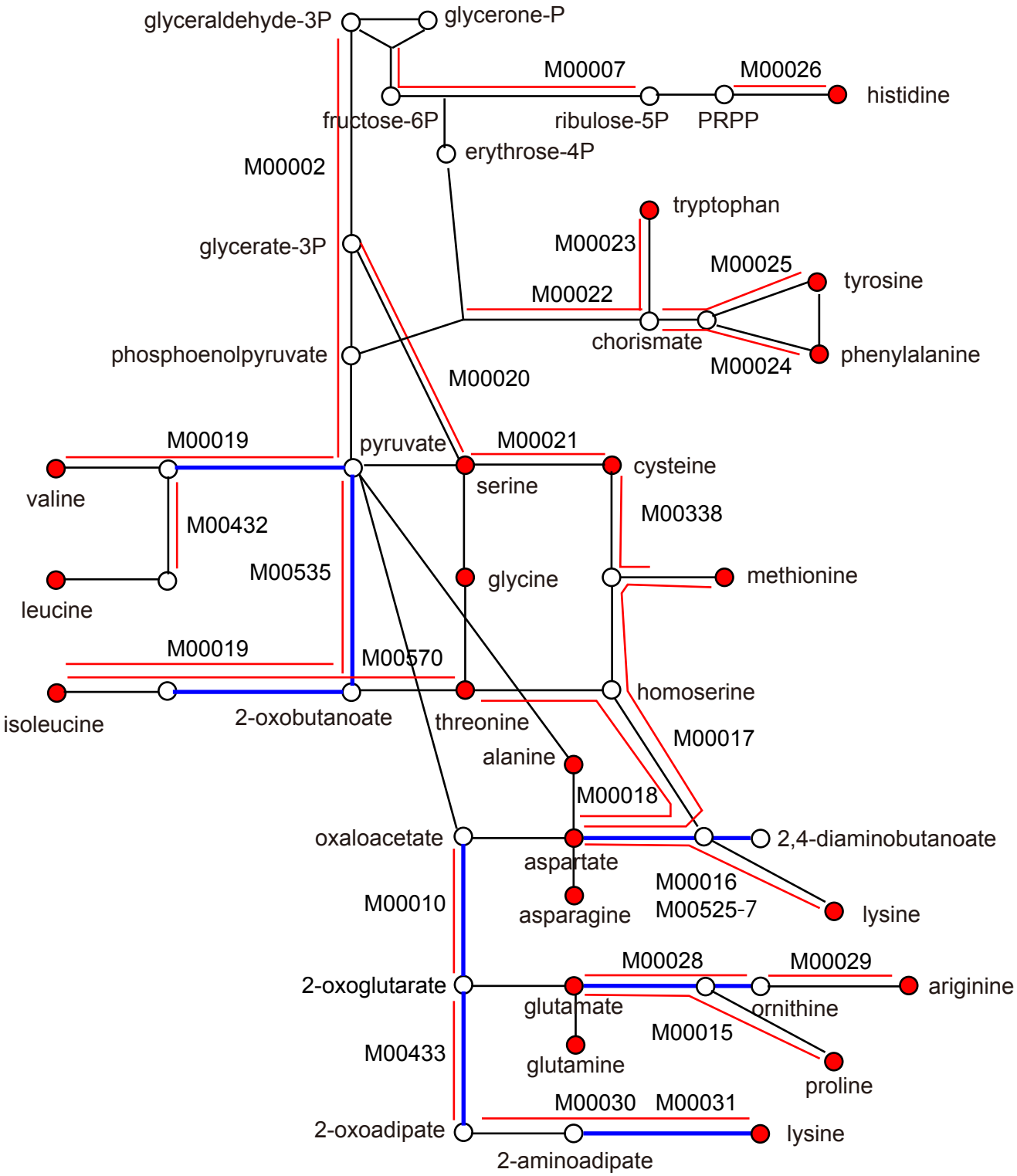

Figure 2

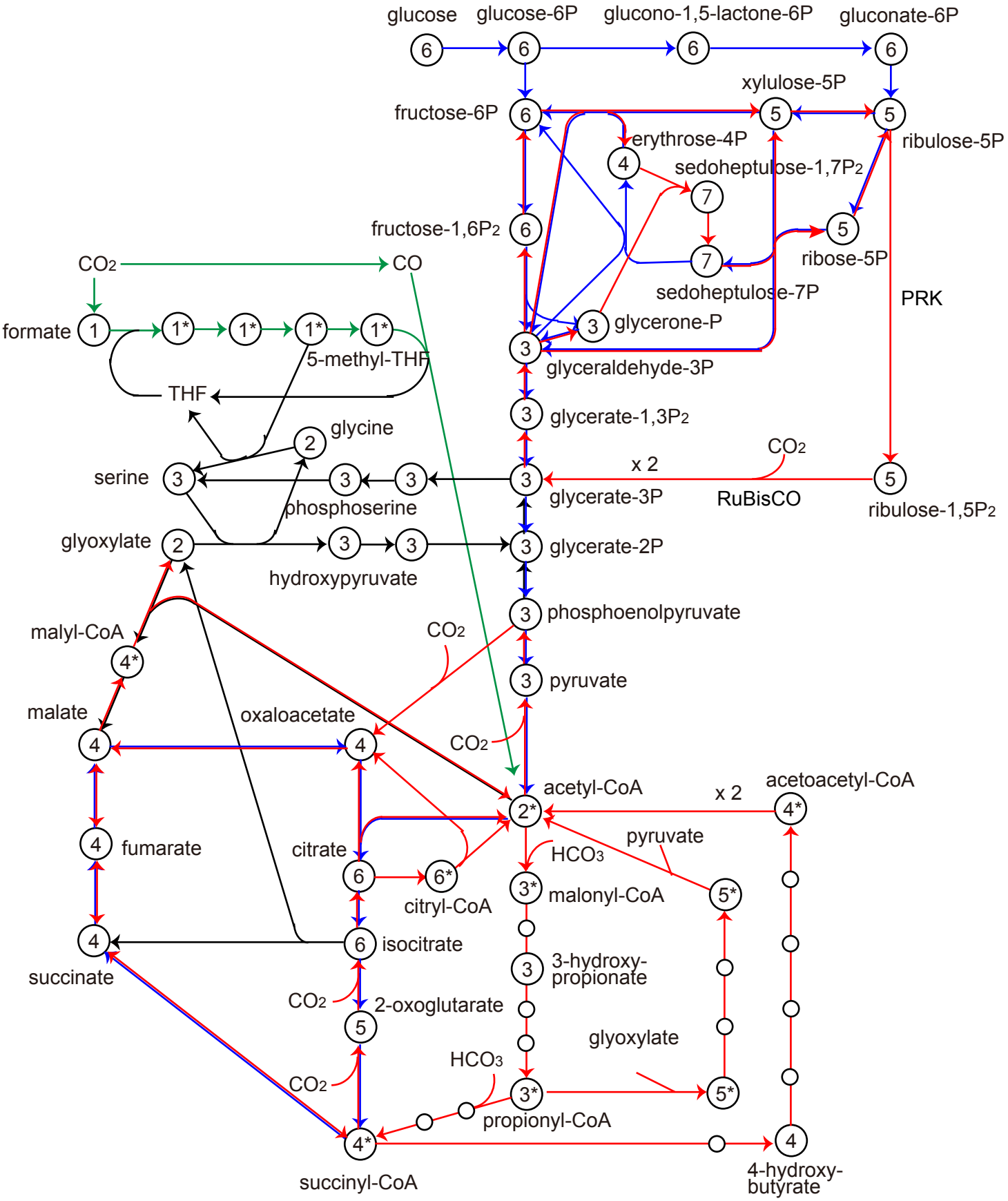

Figure 3